Computational Linguistics and Chinese Language Processing
Vol. 25, No. 1, June 2020, pp. 57-80

應用多模式特徵融合的深度注意力網路進行謠言檢測[1]

Rumor Detection Using Deep Attention Networks With Multimodal Feature Fusion

王正豪*、黃靖幃*

Jenq-Haur Wang and Chin-Wei Huang

摘要

隨著社群平台蓬勃的發展，許多謠言與假訊息也充斥在社群媒體之中。現今各大社群平台大多是透過人工的舉報或統計的方式來進行謠言的分辨，這在資訊快速傳播的時代，非常缺乏效率。本論文提出一個結合圖像描述模型的多模式特徵融合方法，並透過深度注意力網路來進行謠言檢測。從 Tweets 中擷取出圖像、文字內容、與發文者的社群特徵後，首先，我們將圖像輸入圖像描述模型，透過 CNN 與 Seq2Seq 模型產生能描述該圖像的語句；其次，這些語句與文字內容串接，經過 word embedding 編碼後，以 Early 及 Late Fusion 兩種特徵融合方式，進一步結合社群特徵。最後，我們設計了多層 (Multi-layer)及多單元 (Multi-cell)雙向遞迴式神經網路 (BRNN)，並結合注意力機制賦予每個特徵不同的權重，以找出最重要的特徵並進行分類。實驗結果顯示，以 Early Fusion 融合所有特徵，使用基於 GRU 的多單元 (Multi-cell) BRNN 架構能達到最佳效果，F-measure 達 0.89，驗證了本論文所提出謠言檢測方法的有效性，未來將以更大量的資料進行實驗。

[1] 本論文承蒙審查委員提供諸多建議，謹此致謝。
* 國立台北科技大學資訊工程系
Department of Computer Science and Information Engineering, National Taipei University of Technology
E-mail: jhwang@csie.ntut.edu.tw

Abstract

With the rapid growth of information, browsing social media on the Internet is becoming a part of people's daily lives. Social platforms give us the latest information in real time, for example, sharing personal life and commenting on social events. However, with the vigorous development of social platforms, lots of rumors and fake messages are appearing on the Internet. Most of the social platforms use manual reporting or statistics to distinguish rumors, which are very inefficient. In this paper, we propose a multimodal feature fusion approach to rumor detection by combining image captioning model with deep attention networks. First, for images extracted from tweets, we apply Image Caption model to generate captions by Convolutional Neural Networks (CNNs) and Sequence-to-Sequence (Seq2Seq) model. Second, words in captions and text contents from tweets are represented as vectors by word embedding models and combined with social features in tweets with early and late fusion strategies. Finally, we design Multi-layer and Multi-cell Bi-directional Recurrent Neural Networks (BRNNs) with attention mechanism to find word dependency and learn the most important features for classification. From the experimental results, the best F-measure of 0.89 can be obtained for our proposed Multi-cell BRNN based on Gated Recurrent Units (GRUs) with attention using early fusion of all features except for user features. This shows the potential of our proposed approach to rumor detection. Further investigation is needed for data in larger scales.

關鍵詞：謠言檢測、遞迴式神經網路、注意力機制、圖像描述、特徵融合

Keywords: Rumor Detection, Bi-directional Recurrent Neural Networks, Gated Recurrent Unit, Self-attention Mechanism, Multimodal Feature Fusion

1. 緒論 (Introduction)

隨著社群網路快速發展，人們可以即時從各大社群平台獲取最新訊息。然而謠言及假訊息充斥其中，如何分辨訊息的真假，避免人們被誤導，是現今各大社群平台所面臨的重大問題。較著名的社群網站，如 Facebook、Twitter 等，針對謠言的辨別都有相應的處理機制。Facebook 利用公正的第三方機構對訊息進行人工驗證，得知訊息的真偽；而 Twitter 則利用自動評估系統與人工標記，標示具爭議或誤導資訊。然而，第三方驗證與人工標記等方法無法即時進行辨識，並阻止假訊息繼續傳播。因此如何快速又準確的辨識假訊息或謠言，已成為近年來熱門的研究議題。

在社群網路謠言檢測的相關研究中，大致可分為針對發文內容，以及透過分析社群網路的傳播結構兩種方法。首先，社群網路發文內容，包含有: 文字、圖像等，可以透過深度學習方法，例如: 遞迴式神經網路 (Recurrent Neural Networks, RNNs)，或卷積神

經網路 (Convolutional Neural Networks, CNNs)，來進行分類以辨識假訊息，但因為發文內容簡短，如果文件數量不足，其學習效果有限。其次，社群網路使用者之間可能有不同的關係，如: 好友、追隨、等，透過社群網路分析方法只著重在發掘關係結構，容易忽略文件內容所表達的資訊，導致謠言的辨識率不佳。有鑒於此，本論文提出一個結合圖像描述模組的多模式特徵融合方法，並透過深度類神經網路架構搭配注意力機制，以提升謠言偵測的準確率。首先，我們提出了利用圖像描述模型 (Image Captioning Model)，以 CNN 擷取圖像特徵，並透過 Sequence To Sequence (Seq2Seq)概念 (Sutskever, Vinyals & Le, 2014)，將圖像轉換為能夠表達圖像內容的文字。其次，我們設計了多層 (Multi-layer)以及多單元 (Multi-cell) 兩種雙向遞迴式神經網路(Bi-directional RNNs, BRNNs)，結合自注意力機制 (self-attention)，並利用 Early 及 Late Fusion 兩種特徵融合方法結合文字、圖像、及社群特徵，以提升分類準確率。雖然過去相關研究已經有論文提出多模式特徵融合的深度神經網路，結合文字、圖像、及社群等特徵進行謠言檢測，如: Jin 等人的作法 (Jin, Cao, Guo, Zhang & Luo, 2017)，其中文字內容透過 Long Short-Term Memory (LSTM)及注意力機制 (attention)，擷取特徵並計算出 attention 權重；而圖像內容則是直接以 CNN 架構取出特徵，並且將 attention 權重與圖像特徵直接進行 elementwise multiplication。然而，這樣的相乘並沒有具體可解釋的實質意義，因為 CNN 所取出的特徵向量與 LSTM 後 atttention 的特徵向量之間，維度不相同且兩向量各維度並沒有任何關聯。 而且現有方法的深度神經網路架構僅採用 LSTM 及 attention，隨著更多深層的神經網路模型不斷進步，仍有進一步改善的空間。因此在本論文所提出的方法中，我們結合了圖像描述模型，如: Vinyals 等人 (Vinyals, Toshev, Bengio & Erhan, 2015) 和 Xu 等人 (Xu *et al.*, 2015) 所提方法，先將圖像內容轉換成最可能的文字描述，以提升圖像特徵的語意，然後再與其他文字，進行 word embedding 及特徵融合；同時我們設計了多層(Multi-layer)及多單元 (Multi-cell)雙向遞迴式神經網路(Bi-directional RNNs, BRNNs)，結合自注意力機制 (self-attention)，並以 Gated Recurrent Unit (GRU)取代 LSTM，以提升分類效果。本文主要的貢獻為:

(1) 我們是第一個結合圖像描述模型的多模式融合謠言偵測方法，讓圖像內容的融合具有意義。比起現有作法，能有效提升準確率。

(2) 我們提出創新的多單元雙向遞迴式神經網路 (Multi-cell BRNN): 在 forward 及 backward 雙向的 RNN，以多個記憶單元 (memory cells)，同時進行序列資料的記憶與學習，能進一步提升效果。

實驗結果顯示，使用基於 GRU 的多單元雙向遞迴式神經網路 (Multi-cell BRNN)搭配注意力機制，可以使分類結果的 F-measure 達到 0.816；在進一步以 Early Fusion 融合社群特徵後，能達到最佳的謠言檢測率，F-measure 可達 0.89，驗證了本論文所提出方法的效果。後續我們在第二章介紹相關研究，第三章詳述研究方法，第四章則描述實驗結果及分析，第五章則是結論。

2. 相關研究 (Related Work)

隨著社群平台上大量使用者產生內容 (user generated content)的快速出現，謠言檢測已成為不可忽略的議題。不管是在 Facebook 或 Twitter，都提供錯誤資訊的檢測機制，以驗證使用者發文的真實性。Facebook 透過使用者與第三方檢查機構協助，對不實訊息進行標註，被標註的訊息會經過 FactCheck.org 和 Snopes.com 等第三方事實查核機構驗證。若經驗證確認為假訊息，則該訊息將會被公開。Twitter 則透過使用者的標註，以自動評估系統賦予每一則推文可信度等級。若可信度等級過低或該訊息內容被一定程度的用戶標示為假訊息，則判斷該訊息為謠言。然而，在這個資訊傳播快速的時代，第三方驗證與人工標記方法都無法即時辨別假訊息並阻止其繼續傳播。如何快速又準確的進行謠言檢測，即是本論文主要探討的議題。

謠言檢測的特徵來源主要可分為兩大類: 發文內容，以及傳播路徑。透過發文內容特徵擷取，以及發文被分享及再分享等行為，作為謠言檢測的特徵，並以機器學習方法訓練模型。例如: Castillo 等人 (Castillo, Mendoza & Poblete, 2011) 根據 tweets 的文字內容，再加上使用者的發文與 retweet 行為，以及引用外部來源等特徵，以 decision tree 來判斷 Twitter 的資訊可信度 (information credibility)。Gupta 等人 (Gupta, Zhao & Han, 2012) 以類似 PageRank 的方式進行 authority propagation，並且依據相似事件應該有相似可信度的想法，計算出可信度的值。

近年來人工智慧再度受到重視，大多謠言檢測相關論文都使用深度學習方法。例如: Ma 等人 (Ma *et al.*, 2016) 利用 RNN 來檢測 Weibo 與 Twitter 推文是否為謠言； Yu 等人 (Yu, Liu, Wu, Wang & Tan, 2017)和 Chen 等人 (Chen, Li, Yin & Zhang, 2018) 分別提出基於 CNN 的錯誤訊息識別卷積法(CAMI)與深度注意力機制，嘗試在發文早期判斷該推文是否為假訊息；Ma 等人 (Ma, Gao & Wong, 2018) 將立場檢測任務與謠言檢測任務整合，試圖透過判別訊息的立場來輔助假訊息的判斷；Jin 等人 (Jin *et al.*, 2017) 則是結合社群網路上的多媒體訊息，如: 文字、圖像、及社群特徵，其中文字內容透過 LSTM 及注意力機制，擷取特徵並計算出注意力權重；而圖像內容則是直接以 CNN 架構取出特徵，並且將注意力權重與圖像特徵直接進行 elementwise multiplication。然而，這樣的相乘並沒有具體可解釋的實質意義，因為 CNN 所取出的特徵向量與 LSTM 後注意力權重向量之間，維度不相同且兩向量各維度並沒有任何關聯。同時在該論文中，對於文字特徵的處理，僅使用 LSTM 進行特徵擷取，隨著更多深層的神經網路模型不斷進步，仍有進一步改善的空間。因此本論文針對以上兩點問題進行改善：首先，針對圖像特徵部分，我們使用一句短語來表達該圖像的內容，比起直接用 CNN 特徵向量來代表圖像，更能表達該圖像的語意。其次，針對文字特徵部分，我們使用雙向遞迴式神經網路(BRNN)結合注意力機制來取得發文內容字詞之間的關係，並試圖找出重點字詞，使後續謠言分類效果得以提升。

隨著電腦運算能力的提升與圖形處理器（Graphics Processing Unit, GPU）的發展，深度學習方法特別是各種類神經網路架構成為熱門的研究方法。CNN 最初是由 Yann

LeCun 等人提出 (LeCun, Bottou, Bengio & Haffner, 1998)。其概念是透過卷積網路層（Convolutional）與池化網路層（Pooling）使輸入的訊息可以保留更多的特徵，不像基本的神經網路只能取得輸入資料一個維度的特徵。CNN 通常用在處理圖像相關的任務，目前已經有許多不同變異的架構應用在各領域，例如著名的 VGG Net (Simonyan & Zisserman, 2015)與 GoogleLeNet (Szegedy *et al.*, 2015) 等架構都在解決傳統卷積層在特徵傳遞過程中，因為某些特徵不夠明顯而被忽略的問題。RNN 最初是由 Elman 所提出 (Elman, 1990)，後來被 Mikolov 等人 (Mikolov, Karafiát, Burget, Černocký & Khudanpur, 2010) 應用在自然語言處理中。RNN 的主要架構如圖 1 所示，是由單層隱藏層的神經網路不斷遞迴而成的。

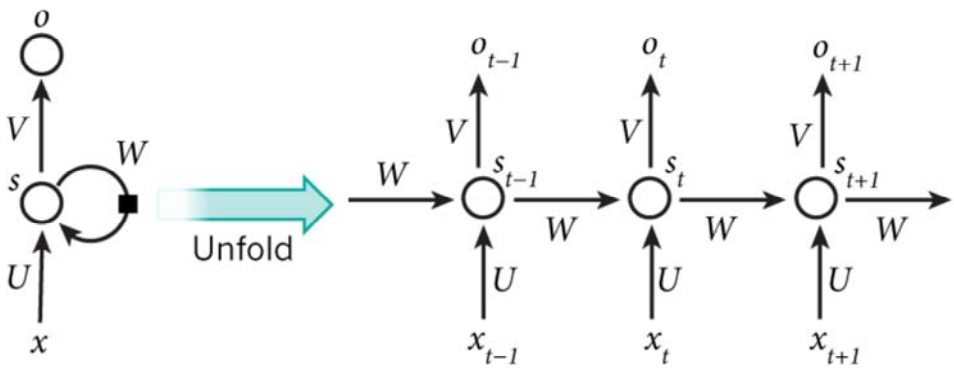
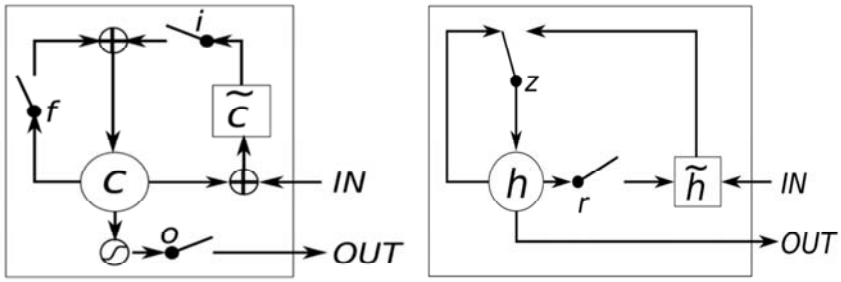


圖 1. 遞迴式類神經網路架構圖 (LeCunn et al., 2015)
[Figure 1. The architecture of recurrent neural networks (LeCunn et al., 2015)]

由上圖可得知，若輸入的資料是一串序列，則資料將會按照時間順序依次輸入至隱藏層，並將上一時間點隱藏層的輸出作為下一時間點隱藏層的輸入。透過這樣的方式，可以使每個時間點的輸出都能與上個時間點的輸入有關，讓神經網路能對整個序列順序進行記憶並學習。

然而，由於 RNN 是以 Back-Propagation Through Time（BPTT）的方式進行訓練與傳遞特徵，容易因為特徵權重的大小，影響下一層隱藏層輸出的資訊，進而導致神經網路可能無法學習到長時間的訊息，其問題稱之為梯度爆炸或梯度消失。目前已經有許多方式來解決該問題，其中最常見的方法就是利用長短期記憶神經網路（LSTM）進行改善。透過三個 Gate：Input Gate、Forget Gate、Output Gate，控制資訊的流動，確保特徵不會因為權重太小而被神經網路忽略。Cho 等人 (Cho *et al.*, 2014) 提出一個嶄新的架構，稱為 Gated Recurrent Unit (GRU)，則進一步簡化處理單元。經過 Chung 等人(Chung , Gulcehre, Cho & Bengio., 2014)的實驗與探討，發現 GRU 其不僅與 LSTM 一樣可以解決遞迴式神經網路的梯度爆炸與梯度消失問題，時間效率也比 LSTM 更好。LSTM 與 GRU 之架構分別如圖 2(a)及 2(b)所示：



(a) *(b)*

圖 2. (a) LSTM Cell 與 (b) GRU Cell 架構圖 (Chung et al., 2014)
[Figure 2. The architectures of an LSTM Cell and a GRU Cell (Chung et al., 2014)]

GRU 透過圖 2 中的 z（update gate）與 r（reset gate）共同控制當前時間點的隱藏狀態。Update gate 負責決定要將多少訊息傳遞到下一個時間點；Reset gate 負責決定要遺忘多少過去的訊息。GRU 現今已經廣為採用，成為解決遞迴式神經網路梯度消失的主流辦法。由於在各種不同的任務中，GRU 與 LSTM 常有不同的表現，而且 GRU 使用較少的 gates，架構簡單效率較佳，因此在本論文中，我們將比較基於 GRU 與 LSTM 的 RNN 架構，對於謠言偵測的效果。

Sequence To Sequence（Seq2Seq）概念最早由 Sutskever 等人所提出 (Sutskever *et al.*, 2014)，被應用在機器翻譯的任務。將輸入的句子（Sequence）經過學習，產生另一個句子（Sequence）。Seq2Seq 架構主要是由兩個遞迴式神經網路所組成，分別稱為 Encoder 與 Decoder，其架構如圖 3 所示:

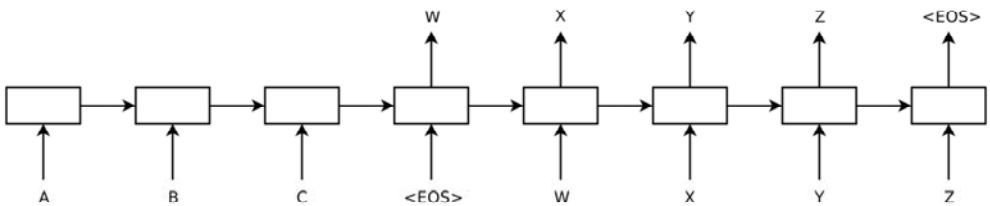


圖 3. Seq2Seq 架構示意圖，輸入”ABC”以產生”WXYZ” (Sutskever et al., 2014)
[Figure 3. The architecture of Seq2Seq Model, which outputs “WXYZ” for input “ABC” (Sutskever et al., 2014)]

在 Encoder 階段，RNN 不斷學習輸入 sequence 中的特徵，在遇到終止符號（<EOS>）時，類神經網路停止編碼並開始 Decoder 的階段，根據前面的記憶，產生一個代表該句子的向量（W），稱之為 context vector，再將它傳入 Decoder，訓練神經網路輸出最接近對應文件的向量，直到出現終止符號。Seq2Seq 架構被應用在許多情境，例如: Facebook 團隊的 Gehring 等人提出 ConvSeq2Seq (Gehring, Auli, Grangier, Yarats & Dauphin, 2017)，將 CNN 與 Seq2Seq 結合，以提升文件翻譯時的速度與準確率；Xing 等人(Xing *et al.*, 2017)

提出一個主題感知的 Seq2Seq 模型，並應用在聊天機器人中，為聊天機器人生成更多訊息豐富且有趣的回應。Zhao 等人 (Zhao *et al.*, 2018) 則基於 Seq2Seq 架構，結合 CNN 的圖像 encoder 與 LSTM 的文字 decoder，進行圖像描述。本論文應用類似想法，產生圖像所對應的文字描述特徵，以進行謠言檢測。

與 RNN 類似的，Seq2Seq 架構也會因為訊息過長而導致梯度消失的問題。雖然 LSTM 常被用來解決該問題，但其效果有限。透過注意力機制 (attention)，可以使神經網路在進行計算時，加強關注與輸入資訊相關的重點特徵，而不只是侷限在經過 RNN 計算後的隱藏向量。Mnih 等人 (Mnih, Heess, Graves & Kavukcuoglu, 2014) 首度將注意力的概念與 RNN 結合，應用在圖像分類任務之中。Bahdanau 等人 (Bahdanau, Cho & Bengio, 2015) 首先將注意力機制用在自然語言處理的任務上。結合雙向 RNN 之注意力機制如圖 4 所示：

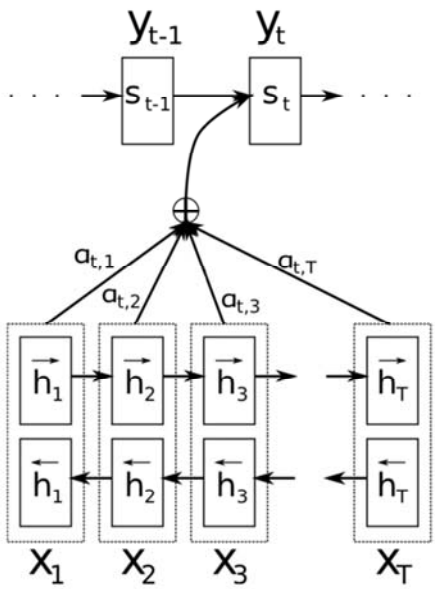


圖 4. 結合雙向 RNN 之注意力機制架構 (Bahdanau et al., 2015)
[Figure 4. The architecture of attention mechanism combining Bidirectional RNNs (Bahdanau et al., 2015)]

如圖 4 所示，輸入文件$X_1, X_2, \dots, X_T$之後，首先，先透過雙向遞迴式神經網路(BRNN)得到各個隱藏層的狀態$h_1, h_2, \dots, h_T$，其中h_j=$\overrightarrow{h_j^T}, \overleftarrow{h_j^T}$。假設當前 Decoder 的狀態為$S_{t-1}$，則輸入與輸出之間的關係可以表示為：

$$\overrightarrow{e_t} = (a(S_{t-1}, h_1), a(S_{t-1}, h_2), \dots, a(S_{t-1}, h_T)) \quad (1)$$

其中a為計算相關性的函數，例如內積或加權內積等。其次，透過 Softmax 函數，對 $\overrightarrow{e_t}$ 進行正規化，即得到注意力權重α_{tj}，定義為：

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})} \quad (2)$$

最後，將注意力權重與各個隱藏層狀態h_j進行加權運算，得出 Encoder 的輸出向量（context vector）$\overrightarrow{c_t}$，並傳入 Decoder 中，其公式為：

$$\overrightarrow{c_t} = \sum_{j=1}^{T} \alpha_{tj} h_j \quad (3)$$

RNN Decoder 的 hidden state 為 S_t，最後輸出為 y_t，S_t 由前一個 hidden state S_{t-1}，前一個輸出 y_{t-1}，以及$\overrightarrow{c_t}$經過函數 f 計算而得。本論文將 GRU 及 LSTM 等 RNN 架構結合注意力機制，以提升神經網路對重要特徵的關注程度，使謠言偵測的準確度得以提升。

3. 研究方法 (The Proposed Method)

本論文提出的方法主要可分為五大步驟，分別為：特徵擷取 (Feature Extraction)、圖像描述 (Image Captioning)、特徵融合 (Feature Fusion)、遞迴式神經網路 (Recurrent Neural Network)、注意力機制 (Attention Layer)，如圖 5 所示。

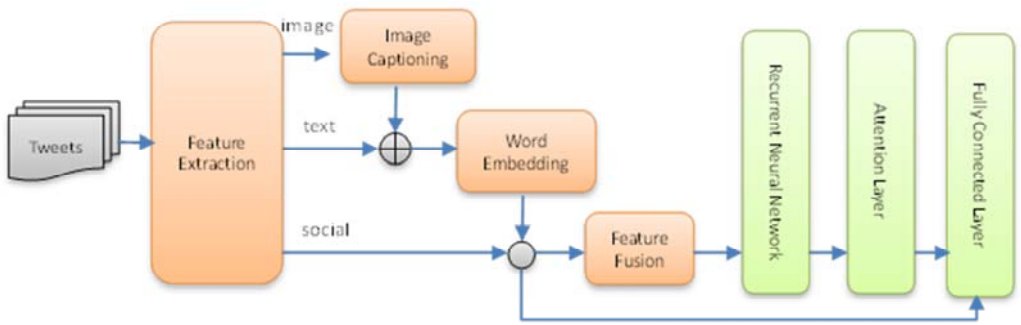


圖 5. 系統架構圖

[Figure 5. The system architecture of the proposed approach]

如圖 5 所示，Twitter 上的推文先經過 Feature Extraction，取得文字內容、圖像、與社群特徵。首先，將圖像特徵輸入圖像描述模組，經過卷積神經網路（Convolution Neural Network）與 Sequence to Sequence (Seq2Seq) 神經網路架構計算後，產生出描述該圖像的語句。其次，語句與文字內容串接，經過 Word Embedding 編碼，透過 Feature Fusion 與社群特徵融合。接著，融合後的特徵向量傳入雙向遞迴式類神經網路層（Bi-directional Recurrent Neural Network, BRNN），找出文字內容中各字詞之間的關係。我們以單層 BRNN 為基礎，設計出兩種不同的堆疊方式，分別為多層雙向遞迴式神經網路（Multi-layer BRNN）、多單元雙向遞迴式神經網路（Multi-cell BRNN）。最後，透過注意力機制（Attention Layer）的計算，加強推文中重要字詞的權重，並輸入一個全連接層（Fully Connected Layer），以進行假訊息的分類。

3.1 特徵擷取 (Feature Extraction)

一篇推文 (Tweet)中通常包含了文字敘述、圖像訊息以及社群資訊。首先，我們擷取出推文中的文字敘述，透過各種 RNN，嘗試找出文字內容的上下文關係。其次，通常推文常含有與文字內容相關的圖像，因此我們利用圖像描述模組，擷取圖像特徵並產生描述該圖像訊息的短句，以找出圖像中隱含的語意。此外，我們考慮各種社群特徵，包括推文的情緒極性、推文中的標籤(hashtag)、及發文者的使用者特徵等。使用者在推文中常會表達個人的意見或情緒，因此我們透過情緒分析模組，採用 SentiWordNet (Esuli & Sebastiani, 2006)對字詞進行情緒極性的擷取，透過計算出推文中每個字詞的情緒分數，加總平均後得出該推文所表達的意見傾向，包括: 正面、中立、負面。社群使用者也常透過 hashtag 標示本文重點主題，對分類可能有幫助。此外，我們也考慮使用者之間的互動作為使用者特徵，包括: 追隨、發文、與回覆等，為了與相關論文進行公平比較，在使用者特徵的部份我們採用與 Jin 等人 (Jin *et al.*, 2017) 相同的特徵，包括: 使用者在 Twitter 的朋友數量、追隨數量、追隨數量中是朋友的比例、總發文數量與是否有被 Twitter 認證等。最後結合使用者特徵、情緒特徵、與標籤特徵便構成社群特徵。

3.2 圖像描述 (Image Captioning)

我們參考 Vinyals 等人所提出的架構 (Vinyals *et al.*, 2015)，使用結合 CNN 與 LSTM 組成的 Seq2Seq 網路架構，產生出能描述該圖像的文字敘述。模型架構如圖 6 所示：

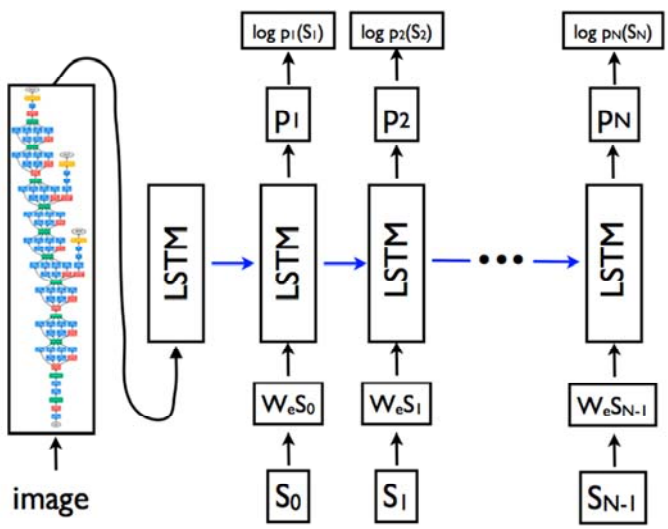


圖 6. 圖像描述模型架構圖 (Vinyals et al., 2015)
[Figure 6. The architecture of image captioning module (Vinyals et al., 2015)]

如圖 6 所示，圖像的部分採用 Google Inception Net V3 的 CNN 架構。此架構共有 42 層，共使用了 4 種不同維度大小的卷積核，可以取得圖像在不同尺度下的特徵，避免

一些細微特徵被忽略。經過 Inception 模組結取出來的圖像特徵向量會與經過 one-hot 編碼的文字敘述一同輸入 LSTM 運算。運算過程中圖像的特徵向量只會輸入一次，之後每個時間點依序輸入文字敘述中的字詞(S_i)，並輸出相關分數 p_i 最高的 k 個候選詞。而這些候選詞會分別輸入到下一個時間點，與當時所輸出的候選詞結合後，從中選出相關分數最高的字詞，再傳入下一個時間點。經過迭代後，最後一個時間點將會輸出一個融合了圖像訊息的完整文字描述。

3.3 特徵融合 (Feature Fusion)

在擷取了三種多模式的特徵，包括: 文字特徵、圖像特徵與社群特徵之後，我們提出特徵融合的方法來整合不同特徵。推文的文字特徵採取 one-hot 編碼的方法，用 300 維的向量來表示每個字詞的特徵。圖像在經過圖像描述模型轉換成描述語句後，將該語句轉換成與推文的文字特徵同樣 300 維的向量。我們也從推文的文字訊息中擷取出標籤(hashtag）與該推文所表達的情緒。在推文情緒方面，我們根據擷取的情緒分數，分為三種類別：正面、中立、負面。若該推文的情緒分數總分為大於 1，則它視為正面；若情緒分數總分小於 0，則視為負面；若情緒分數總分介於 0 到 1 之間，則視為中立。最後將上述的文字特徵、圖像特徵、以及情緒與 hashtag 兩者經過 one-hot 編碼後的向量進行串聯(concatenate)，即得到所有特徵的向量。

由於社群特徵與其他特徵差異很大，我們考慮兩種不同的特徵融合策略: 早期融合(early fusion)，和晚期融合 (late fusion)。在早期融合的策略中，我們同樣利用 one-hot 編碼，將社群特徵轉換成向量。為了讓社群特徵與圖文特徵能有相當的重要性，我們利用一個 autoencoder，將社群特徵壓縮為 300 維，並且串接在圖文特徵之後，以訓練分類器。而在晚期融合的策略中，我們先以圖文特徵輸入 RNN 和注意力機制，得到一系列的輸出，然後再將社群特徵以 one-hot 編碼轉換成向量，與圖文輸出結果一起輸入 Fully Connected Layer 進行分類。

3.4 遞迴式神經網路 (Recurrent Neural Networks)

本論文在 RNN 模組中使用 GRU Cell 取代傳統的 LSTM，並設計多層的 BRNN 堆疊的架構，以探討謠言偵測的效果。

單層雙向遞迴式神經網路（Bi-directional Recurrent Neural Networks, or BRNNs）最早是由 Schuster 等人提出 (Schuster & Paliwal, 1997)，分別將遞迴神經網路中每一個訓練序列分成向前傳遞（forward pass）與向後傳遞（backward pass）。兩者分別是獨立的單向 RNN，且兩個神經網路都連接到同一層輸出層，如圖 7 所示：

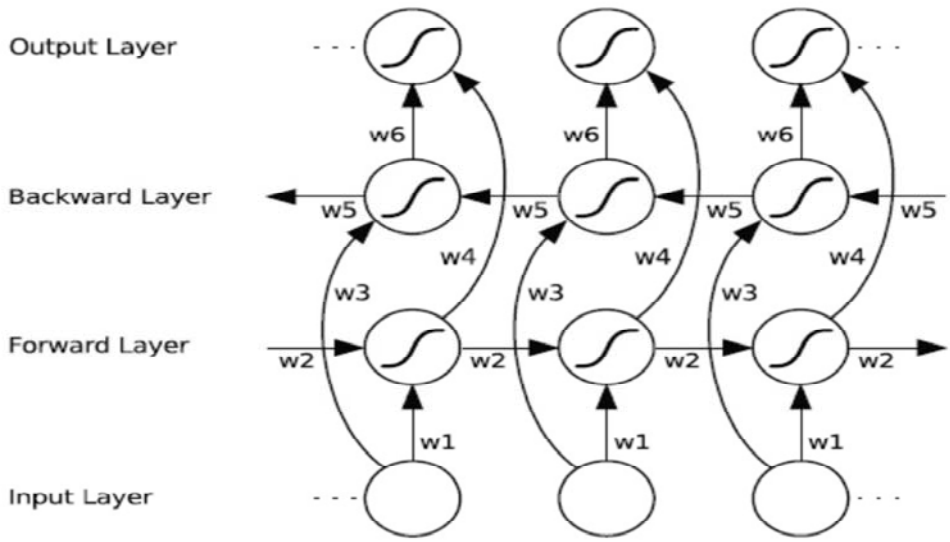


圖 7. 雙向遞迴式網路架構圖 (Graves, 2012)
[Figure 7. The architecture of bi-directional recurrent neural networks (Graves, 2012)]

對於雙向遞迴式神經網路的結果，我們透過串接的方式來組合向前與向後傳遞的輸出以表示一個字詞。

基於單層的雙向遞迴式神經網路，本論文設計了以下兩種不同的方式，進行多層雙向網路的堆疊。首先，多層雙向遞迴神經網路 (Mutli-layer BRNN)，是透過讓文字訊息經過多個回合的雙向遞迴式網路計算，強化文件中字詞之間的相互關係。架構如圖 8 所示：

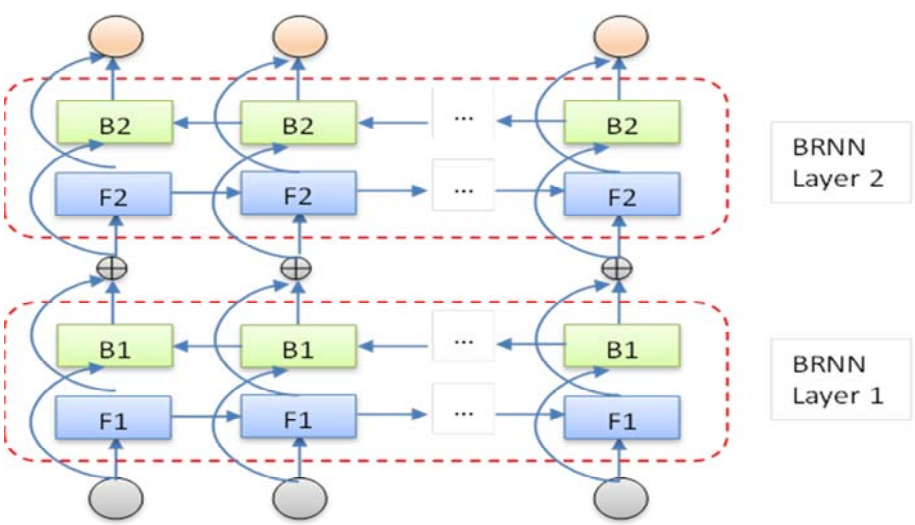


圖 8. 多層雙向遞迴式神經網路架構圖
[Figure 8. The architecture of Multi-layer BRNNs]

多層雙向遞迴式神經網路的輸入與輸出與單層 BRNN 一樣，透過在輸入層依序輸入文件中的字詞，得出代表該字詞的數值向量。其中，當每個字詞經過第一層的 BRNN 後，其得到的輸出已經包含了文件中向前傳遞與向後傳遞的資訊，因此當第一層 BRNN 的輸出傳入第二層時，不僅保留了原始文件中字詞之間的關係，每個字詞向量也記憶了經過第一層計算後的特徵。當下一層網路在計算時，由於其字詞之間的關係已經在上一層被找出，故不必再重新計算，使得網路能更快速的收斂，提升整體模型的計算效率。

其次，我們設計了另一種堆疊雙向網路的方法: 多單元雙向遞迴式神經網路 (Multi-cell BRNN)，透過增加 BRNN 中每個方向的單元數量，進行更深入的計算，當前 Cell 的輸出會作為下一層 Cell 的輸入，同一個神經元中的多個 Cell 同時進行序列資料的記憶與學習。架構如圖 9 所示：

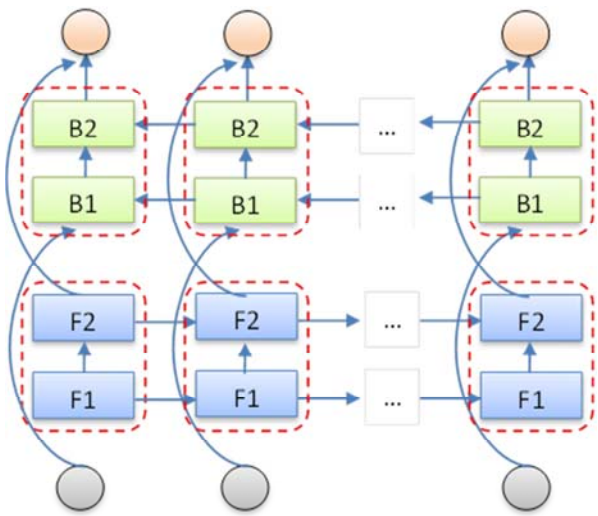


圖 9. 多單元雙向遞迴式神經網路架構圖
[Figure 9. The architecture of Multi-cell BRNNs]

多單元雙向遞迴式神經網路架構，在輸入層依序輸入文件中的字詞，並在輸出層得出代表該字詞的數值向量，如同上述的兩種架構。但由圖 8 可知，對於向前傳遞與向後傳遞而言，每個方向同一時間點的輸入經過一個 Cell 計算後，會再傳入下一個 Cell 繼續計算。該架構與多層雙向遞迴式神經網路不同的是，每個時間點的輸入只有考慮到單一方向的影響，向前傳遞與向後傳遞相對於整體架構來說，還是兩個獨立的神經網路架構，直到最後輸出時才進行整合。此方法能比多層 BRNN 更深入的計算每個字詞之間的關係，但相對的需要花費更多的資源與時間才會讓神經網路收斂。

3.5 注意力機制 (Attention Layer)

注意力機制常被用在遞迴式神經網路，加強關注與輸入資訊相關的重要特徵。在上述的各種神經網路架構中，我們結合自注意力機制 （Self-Attention）來計算文中每個字詞之

間的關係，以取得每則 Tweet 中文字特徵的重要資訊。Self-Attention 是一種注意力機制，與傳統的 Attention 機制差別在於，Self-Attention 不需要透過引入外部的資訊來找出較為重要的訊息，僅需要通過自身的訊息就能更新權重與參數，找出較重要的資訊。它的核心概念是 scaled dot-product attention 架構，是一種 dot-product attention 的變形，如圖 10 所示。

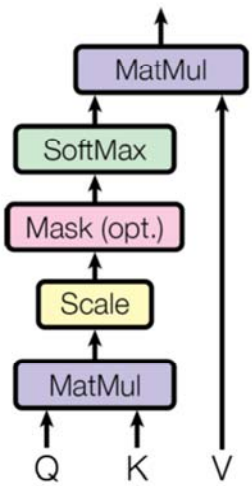


圖 10. Scaled Dot-Product Attention 示意圖 (Vaswani et al., 2017)
[Figure 10. Scaled Dot-Product Attention (Vaswani et al., 2017)]

經過 Vaswani 等人 (Vaswani *et al.*, 2017) 與 Tan 等人 (Tan, Wang, Xie, Chen & Shi, 2018) 的探討與比較，已證實該內積（乘法）注意力機制比使用單層神經網路的標準注意力機制 (Bahdanau *et al.*, 2015) 更有效率。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

其中 d_k 為 key 的維度。當維度越大，Q 與 K 的內積也會越大，因此除以一個調整數 $\sqrt{d_k}$ ，防止該數值結果過大，最後透過 softmax 函數將結果正規化，將獲得的權重與 V 相乘，更新其向量的數值。

為了加速運算，我們採用 Multi-Head Attention 架構，如圖 11 所示：

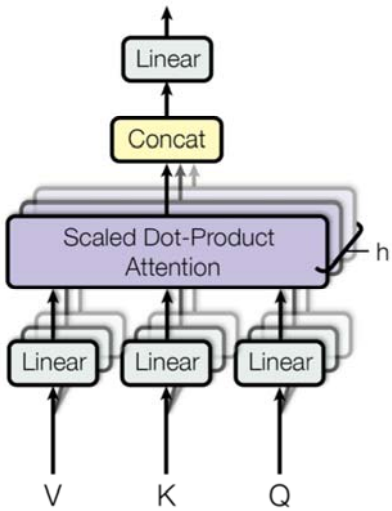


圖 11. Multi-Head Attention (Vaswani et al., 2017)
[Figure 11. Multi-Head Attention (Vaswani et al., 2017)]

經過 h 個不同投影線性轉換後，可以將 h 個 scaled dot-product attention 神經網路進行平行運算，並將每一次的結果進行串接，最後再經過一層線性轉換得到 multi-head attention 的結果。如下所示：

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O$$
$$where\ head_i = Attention\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right) \quad (5)$$

其中 $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$。$i$ 表示為第幾個 scaled dot-product attention 網路，W_i^Q, W_i^K, W_i^V 皆為經過訓練的權重矩陣，W^O 為線性轉換層的權重矩陣，而 $d_k = d_v = d_{model} / h$ 則表示為矩陣的維度。

因為 self-attention 機制是對每個輸入的字詞與所有字詞進行計算，學習文件內部的結構與字詞之間的依賴關係，故計算每個字詞的最大路徑長為 1，即每個字詞都會被計算一次，不會像遞迴式神經網路一樣發生因訊息傳遞路徑過長導致過小的特徵被忽略，進而產生的梯度消失或梯度爆炸問題。在本論文架構中，input 經過各種 RNN，以及 self-attention 機制後，會計算出一連串的 output 向量，我們再將所有向量輸入到全連接層 (Fully Connected Layer)，進行最後的分類。

4. 實驗與討論 (Experiments and Discussions)

本論文採用的資料集分為兩大部份：圖像描述資料集與謠言檢測資料集。首先，在圖像描述方面，我們使用 Microsoft COCO 2014 (Tan *et al.*, 2018) 資料集，它在圖像相關任務中廣泛被使用，包含圖像識別與圖像特徵點檢測，如: Vinyals 等人 (Vinyals et al, 2015) 與 Xu 等人 (Xu *et al.*, 2015)。該資料集的每一張圖像資料都含有 5 句短語進行描述，每一句短語在資料集中皆為唯一。

其次，在謠言檢測方面，由於資料取得不易，每筆訊息都需要經過第三方機構公開認證其真偽，才能確定該訊息是屬於謠言或事實。本實驗採用 MediaEval 2015、2016 任務中所提供的 Twitter 謠言檢測資料集，已經過 Twitter 官方認證，其中也包含了每則推文的多媒體訊息與發文者的相關特徵。兩個資料集的分布狀況如表 1 與表 2 所示：

表 1. 圖像標註資料分布統計
[Table 1. Data distribution in image captioning dataset]

資料集	圖像數量/描述短句數量
Training Data	82783 / 413915
Test Data	36454 / 182270.

表 2. 謠言資料集資料分布統計
[Table 2. Data distribution in rumor dataset]

資料集	推文數量 (event)
Training Data	Real: 189 / fake: 157
Test Data	Real: 21 / fake: 24

在圖像描述的相關實驗中，我們採用雙語互譯評估（Bilingual Evaluation Understudy, BLEU）評估方法來評量圖像描述模型，BLEU 最早是由 IBM 的 Papineni 等人所提出的 (Papineni, Roukos, Ward & Zhu, 2002)，主要是用來評價模型的翻譯結果與參考文件是否相似。BLEU 的定義為: modified n-gram precision 的幾何平均 (geometric mean)，

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} W_n log p_n\right)$$

$$BP = \begin{cases} 1 & if\ \ c > r \\ e^{1-r/c} & if\ c \le r \end{cases} \tag{6}$$

其中c表示譯文的長度，r表示參考文件的長度。

Modified n-gram precision 為 clipped n-gram 個數除以所有 n-gram 個數，

$$P_n = \frac{\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} Count_{clip}(n-gram')} \tag{7}$$

其中 clipped n-gram 個數計算方式如下:

$$Count_{clip} = \min(Count, Max\_Ref\_Count) \tag{8}$$

而在謠言檢測相關實驗中，主要著重在二元分類任務，因此採用準確率（Accuracy）、精確率（Precision）、查全率（Recall）與 F-Measure 進行評估，並利用 T 檢定來比較不

同模型的差異程度。

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

在以下謠言檢測的實驗中，我們主要的 baseline 比較對象皆為 Jin 等人所提出的方法 (Jin *et al.*, 2017)。

4.1 圖像描述的效果 (The Effects of Image Captioning)

首先，為了探討圖像描述模型的效果，針對 MSCOCO 2014 的每一筆圖像資料經過圖像描述模型後，我們利用訓練資料集中的 3 句短文進行模型的訓練，其他 2 句短文進行驗證。為了實驗對照，我們也使用 Mediaeval 2015, 2016 中的圖像訊息進行訓練，BLEU 的實驗評估結果，如圖 12 所示:

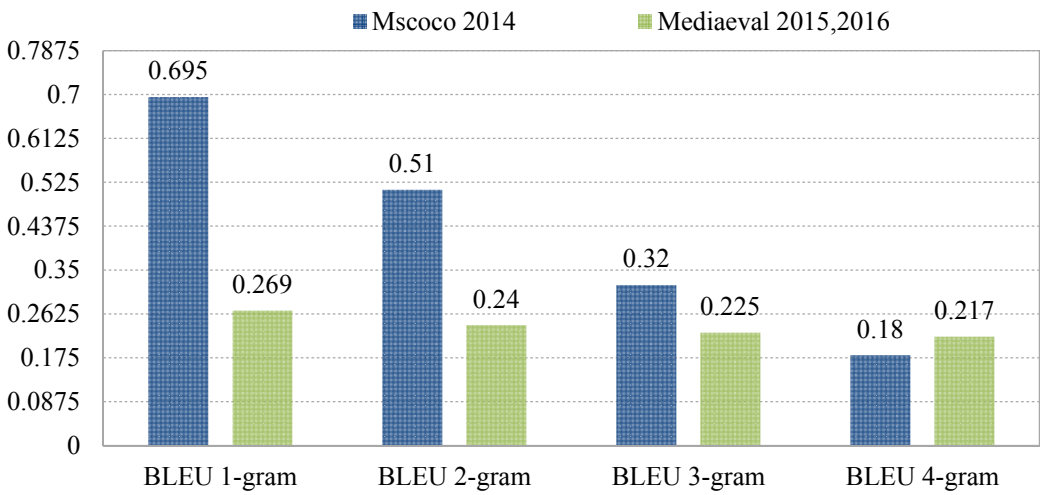


圖 12. 圖像描述實驗評估結果
[Figure 12. Experimental results for image captioning]

如圖 12 所示，以 MSCOCO 2014 資料集訓練的翻譯模型其 BLEU-1, BLEU-2 評估分數分別為 0.695 和 0.51，明顯優於 MediaEval 2015, 2016 所訓練的模型，且相當接近 Xu 等人的結果 (Xu *et al.*, 2015)。由於 Mediaeval 2015, 2016 資料集屬於 Tweets，礙於 Twitter 的資料特性，Tweets 中的文字訊息並不一定是在描述其中的圖像資訊，且發文者與回文者也不一定是客觀的對圖像訊息進行描述，這會使得參考文件不完整，無法訓練出良好的翻譯模型，因此導致最後其 BLEU 分數都偏低，故後續實驗將採用 MSCOCO 2014 資料集訓練出翻譯模型做為謠言檢測的圖像描述模組。

4.2 Word Embedding的效果 (The Effects of Word Embedding)

為了探討文字訊息的向量編碼方法對於謠言偵測的效果，我們使用兩種常見的方法進行比較：隨機初始化法，以及預訓練好的 Word2Vec 字典。前者是從 -1 ~ 1 之間隨機產生代表該字詞的數值向量，之後經過神經網路的訓練進行調整；後者則採用 GoogleNews 預訓練的 Word2Vec 字典對應的字詞向量進行訓練並更新。實驗結果如圖 13 所示：

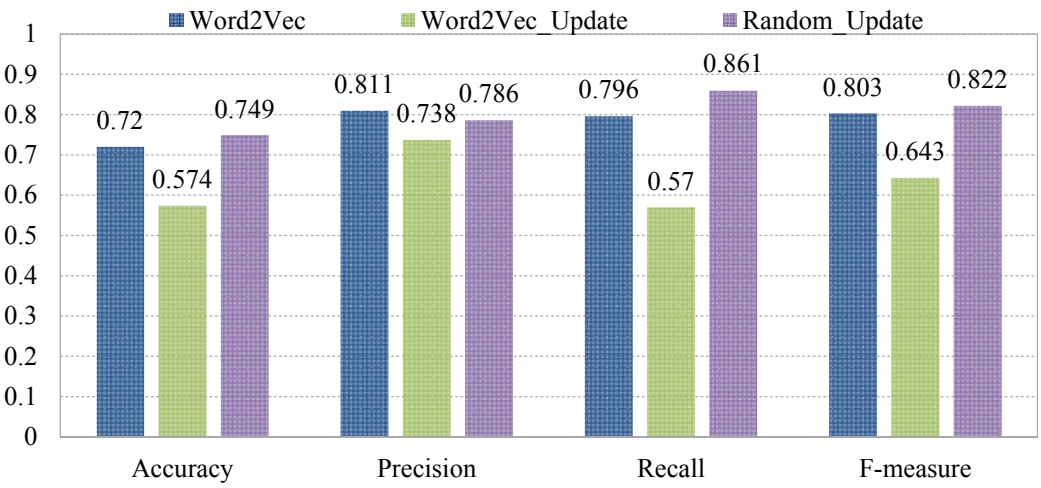


圖 13. Embedding Layer 實驗結果
[Figure 13. Experimental results for embedding layer]

由圖 13 所示，本論文所提方法針對謠言檢測資料集進行訓練時，利用隨機初始化法產生字詞向量會有較好的結果，其 F-measure 高達 0.822。經過多次實驗與探討，發現主要由兩個因素影響此實驗結果。第一，使用 Google News 預訓練的 Word2Vec 字典，每個字詞向量是由許多新聞文章訓練產生，在整個向量空間中彼此都有關聯。若在訓練時使用字典中的字詞向量，並不斷的在 RNN 中更新字詞向量，會使得該字詞在向量空間中的意義被改變，失去與其他字詞的關係，導致最後模型的精準度下降。第二，在謠言檢測資料集中，有一些字詞並未出現在 Google News 的 Word2Vec 字典裡，其字詞像向量為零，導致該字詞被神經網路忽略，進而降低模型的準確率。透過圖 13，我們也發現，使用 Word2Vec 字典但不更新字詞向量的效果較佳，也驗證了預訓練字典裡，若以不同資料來訓練並更新字詞向量，將會失去其原本的意義。

4.3 遞迴式神經網路架構比較 (The Effects of Recurrent Neural Networks)

在此實驗中，我們先僅以文字特徵進行不同 RNN 架構之謠言偵測效果比較，如下圖所示:

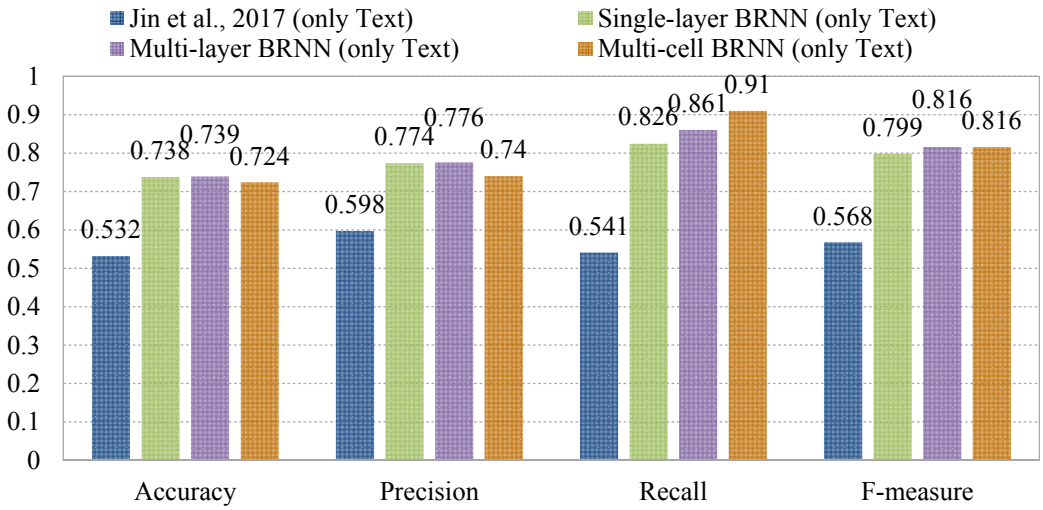


圖 14. 不同遞迴式神經網路架構之比較（文字特徵）

[Figure 14. Experimental results for different recurrent neural networks (Text feature)]

由圖 14 所示，當只考慮文字特徵時，多層 BRNN 與多單元 BRNN 架構的 F-measure 都達到 0.816，效果皆優於 Jin 等人所提出的方法 (Jin *et al.*, 2017)。而由圖 14 也可以得知，在這三種不同的 BRNN 架構中，若只看文字特徵，三者並沒有明顯的優劣，F-measure 都接近 0.8。

4.3.1 特徵融合的效果 (The Effects of Feature Fusion)

接著我們探討結合文字、圖像、與社群特徵，對分類結果的影響。由於社群特徵包含了標籤 (hashtag)、情緒、及使用者，我們首先比較不同組合特徵選取的實驗結果，如圖 15 與圖 16 所示:

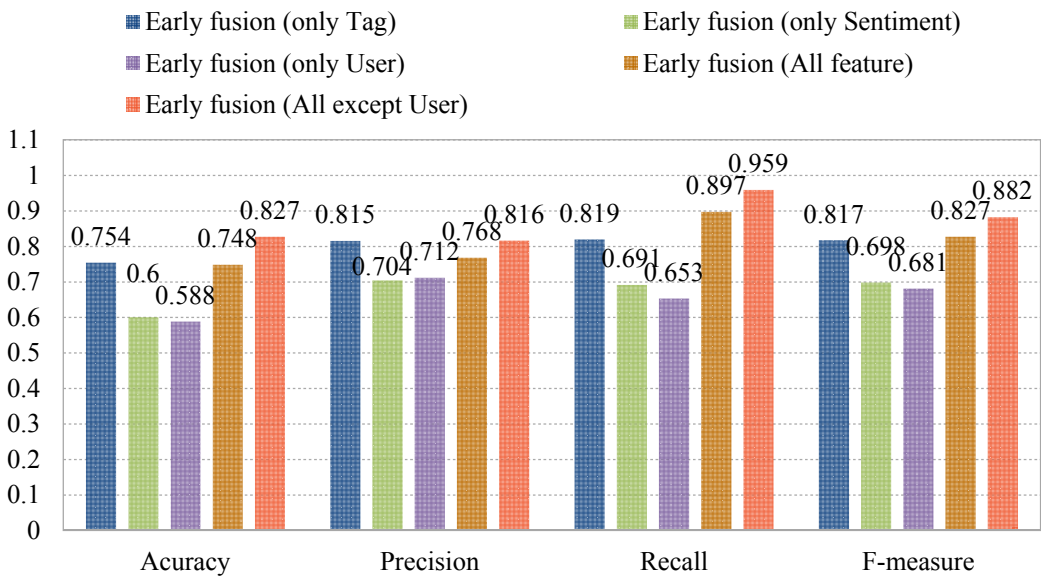


圖 15. Early Fusion 之特徵選取效果比較
[Figure 15. Experimental results for early fusion]

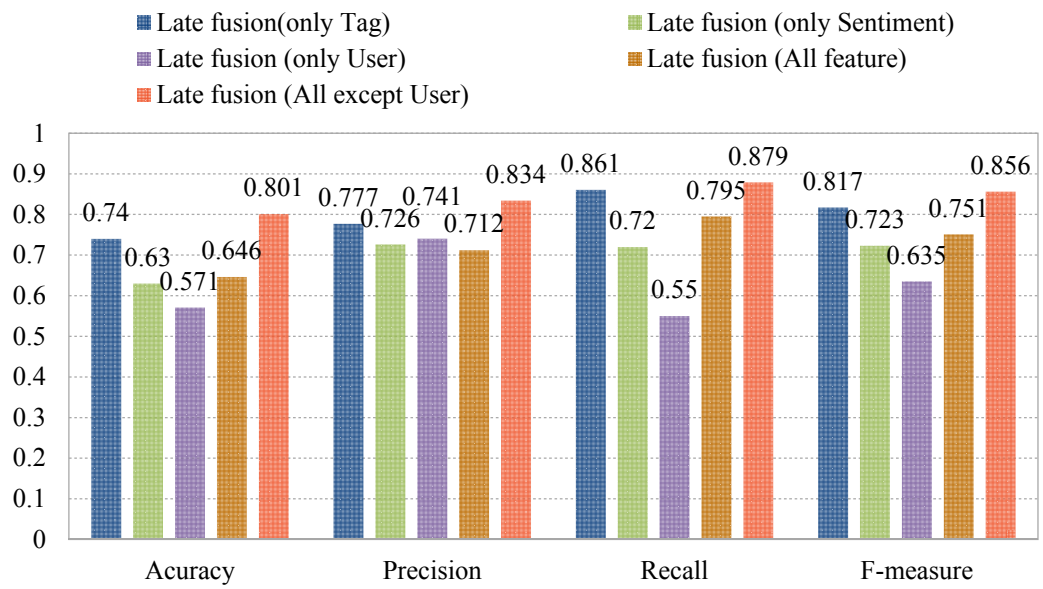


圖 16. Late Fusion 之社群特徵選取效果比較
[Figure 16. Experimental results for late fusion]

由圖 15 與圖 16 所示，不論使用 Early Fusion 或 Late Fusion 特徵融合策略，比起情緒與使用者特徵，加入標籤 (hashtag) 特徵具有較佳的效果。實驗結果顯示加入使用者

特徵反而會降低分類效果，在不採用使用者特徵的情況下，使用 Early Fusion 特徵融合策略，F-measure 最高分別可以達到 0.882 及 0.856，比納入全部特徵時的效果分別提升了 5.5%及 10%。經過觀察，我們發現使用者特徵對於該推文是否為謠言並沒有太大的關係，因為並不會因為該使用者朋友數量或總發文數的多寡，而影響其發出的推文為謠言或事實。

4.3.2 不同RNN架構的效果 (The Effects of RNN Architectures)

在使用 Early Fusion 特徵融合的情形下，我們進行不同 RNN 架構之實驗比較。實驗結果如圖 17 所示:

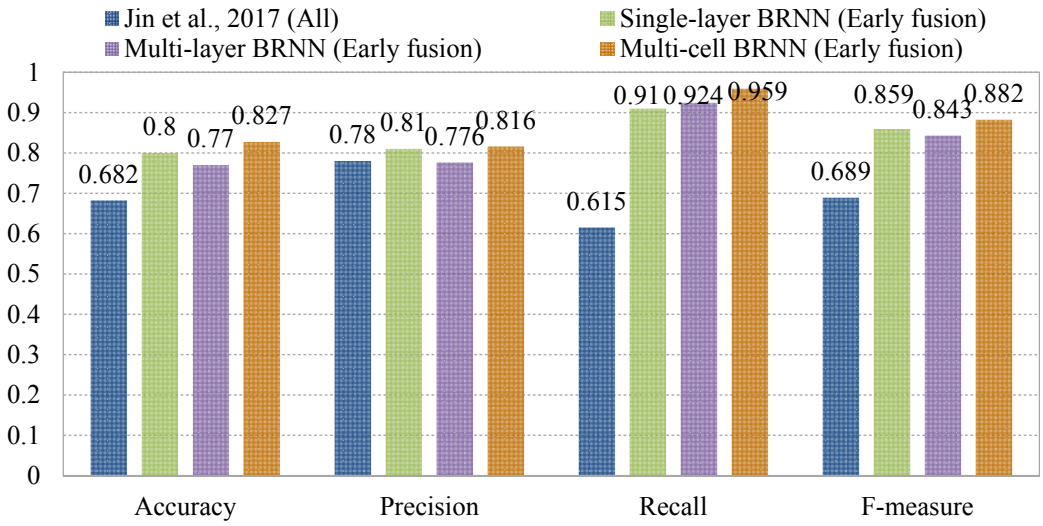


圖 17.不同遞迴式神經網路架構之比較（除使用者之外的所有特徵）

[Figure 17. Experimental results for different recurrent neural networks (All features except user feature)]

如圖 17 所示，在結合除了使用者之外所有特徵後，多單元 BRNN 在所有的評估標準下最為突出，F-measure 達到 0.882，其次為單層與多層 BRNN。經過反覆實驗與探討，我們發現使用多層 BRNN 時，經過第一層 BRNN 後，特徵向量已經過內部神經元運算，找出所有字詞的關聯，並進行了調整，故其輸出的特徵向量已經與原先的輸入不同。且由於該向量與所有特徵的關聯性已經被確定，後續再進入下一層 BRNN 時，該特徵向量並不會再有太大的變動，所以使用單層 BRNN 才會與使用多層 BRNN 的結果相近。為了驗證推論的正確性，本實驗進一步採用 T 檢定，分別針對 F-measure 及 accuracy，進行兩種架構的評比，判斷其差異是否為常態。經過計算，兩種架構針對 F-measure 及 accuracy 的 p-value 分數皆為 0.006，皆小於 0.01，證實其實驗結果並非偶然，具有統計意義。

實驗中我們也發現，多單元 BRNN 雖然類似於單層 BRNN，但是其向前傳遞與向後傳遞神經網路，分別經過了多層的隱藏層計算，透過加深內部 cell 的數目，強化了每個特徵向量與特徵序列（向前傳遞與向後傳遞）之間的關係。最後的實驗結果也顯示了多

單元 BRNN 比其他方法效果好，F-measure 可以達到最高的 0.882。我們也採用多單元 BRNN 與多層 BRNN 進行 T 檢定，針對 F-measure 及 accuracy，兩者的 p-value 數值為 0.04 及 0.014，皆小於 0.05，證實該結果具有統計意義。

接著我們比較兩者特徵融合策略，對最後分類結果的影響。實驗結果如圖 18 所示：

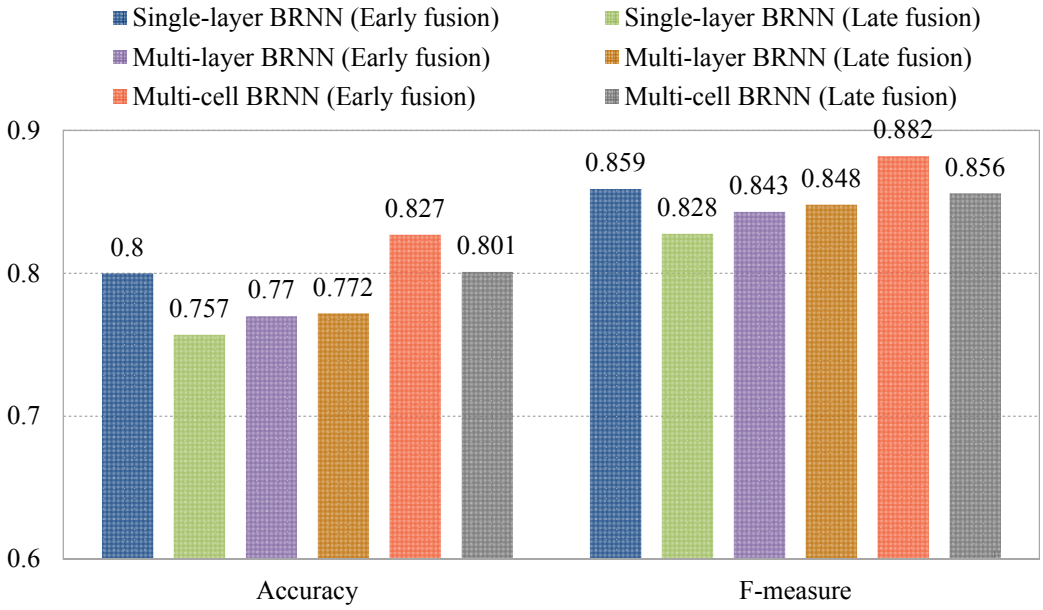


圖 18.不同特徵融合策略之比較

[Figure 18. Experimental results of different feature fusion strategies]

如圖 18 所示，在使用 Late Fusion 特徵融合策略時，不論是多層或多單元 BRNN，兩者的效果非常相近。而 Early Fusion 特徵融合策略對於單層以及多單元 BRNN，能大幅提升其效果。經過多次實驗與探討，我們發現其原因是，透過 BRNN 與注意力機制，社群特徵在 Early Fusion 階段融合，也能像其他特徵一起進行向量權重的調整，找出其中最能代表謠言或非謠言的重要特徵，提升分類準確率。

4.4 LSTM與GRU的比較 (The Effects of LSTM vs. GRU)

最後我們探討 RNN 架構中，採用 LSTM 或 GRU 不同處理單元對分類結果的影響。由於謠言檢測資料集資料量較小，為了降低資料分布的影響，我們採用 5-fold cross-validation。同時由於原 MediaEval 資料集的組成是根據不同事件 (event) 區分為 real 或 fake，並將該事件相關發文與回應 tweets 跟著列為 real 或 fake，並且已經區分為 training 與 test data，因此在進行 5-fold cross-validation 的資料 partition 時，我們將 training 及 test data 分別隨機切為 5 份，分別取 4 份 training 及一份 test，進行 5 次實驗後取其平均。實驗結果如圖 19 所示:

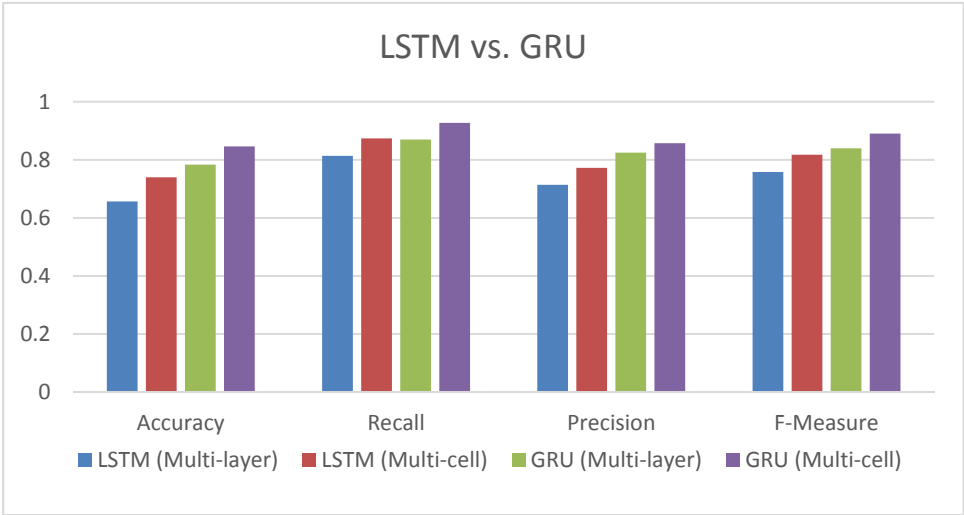


圖 19. LSTM 與 GRU 之比較
[Figure 19. Experimental results of LSTM and GRU]

如圖 19 所示，不論是 accuracy 或 F-Measure，基於 GRU 的 BRNN 架構 (BiGRU) 都比基於 LSTM 的 BRNN 架構 (BiLSTM)效果要好。同時，多單元(Multi-cell) BRNN 也都比多層(Multi-layer) BRNN 效果要好，最佳的效果為 Multi-cell BiGRU，F-Measure 達到 0.89。因此，這也驗證了 LSTM 與 GRU 並沒有絕對的優劣，不同任務須採用不同架構的設計才能獲得較佳的效果。

5. 結論 (Conclusions)

本論文提出一個基於多模式特徵融合的深度類神經網路架構進行謠言檢測。透過圖像描述模型能將圖像轉為敘述文字，有效發掘圖像中的語意，並與文字內容串接，進行 Word Embedding；針對社群特徵，我們採取了 Early 及 Late Fusion 不同特徵融合策略。我們也設計了多層與多單元兩種不同的雙向遞迴式神經網路(BRNN)架構，並結合注意力機制，以提升分類效果。實驗結果顯示，使用基於 GRU 的多單元(Multi-cell) BRNN 架構 (Multi-cell BiGRU)，以 Early Fusion 方式融合社群特徵，並結合文字特徵、圖像描述模組，能有效提升謠言檢測效果, 最佳 F-measure 達 0.89。

本論文所提出的方法仍有限制。首先，我們的方法主要是針對 Twitter 社群平台上的推文進行謠言檢測，由於每則推文的長度並不會太長，所以通常不能直接適用於長文件的謠言檢測。其次，本論文擷取圖像特徵的方式是將圖像先經過圖像描述模型，轉換成有意義的文字訊息。雖然經過評估該模型有一定的精確度，結果經過比對也大致符合圖像所表達的意義，但部分結果還是有落差，因此如何提升圖像描述模型的效果，有待進一步探討。最後，社群特徵並非全部有助於提升分類效果，其中的情緒分類僅採用情緒字典的比對，未來將採取不同的情緒分析方法，以進一步提升謠言偵測的效果。

參考文獻(References)

Bahdanau, D., Cho, K.H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of WWW 2011*, 675-684. doi: 10.1145/1963405.1963500

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining 2018*, 40-52. doi: 10.1007/978-3-030-04503-6_4

Cho, K., Merriënboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., … Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734. doi: 10.3115/v1/D14-1179

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of NIPS 2014*.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211. doi: 10.1016/0364-0213(90)90002-E

Esuli, A. & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 417-422.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning 2017*, 1243-1252.

Graves, A. (2012). *Supervised sequence labelling with Recurrent Neural Networks.* (p.26). German, Heidelberg: Springer.

Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on Twitter. In *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, 153-164. doi: 10.1137/1.9781611972825.14

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia 2017*, 795-816. doi: 10.1145/3123266.3123454

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444. doi: 10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. doi: 10.1109/5.726791

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., … Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of IJCAI 2016*, 3818-3824.

Ma, J., Gao, W., & Wong, K.-F. (2018). Detect rumor and stance jointly by neural multi-task learning. In *Proceedings of The Web Conference 2018*, 585-593. doi: 10.1145/3184558.3188729

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH 2010*, 1045-1048.

Mnih, V., Heess, N., Graves, A. & Kavukcuoglu, K. (2014). Recurrent models of visual attention. In *Proceedings of neural information processing systems 2014*, 2204-2212.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics 2002*, 311-318. doi: 10.3115/1073083.1073135

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681. doi: 10.1109/78.650093

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR 2015*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014) Sequence to sequence learning with neural networks. In *Proceedings of neural information processing systems 2014*, 3104-3112.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015*, 1-9. doi: 10.1109/CVPR.2015.7298594

Tan, Z., Wang, M., Xie, J., Chen, Y., & Shi, X. (2018). Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence 2018*, 4929-4936.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., …Polosukhin, I. (2017). Attention is all you need. In *Proceedings of neural information processing systems 2017*, 5998-6008.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015*, 3156-3164. doi: 10.1109/CVPR.2015.7298935

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., … Ma, W.-Y. (2017). Topic aware neural response generation. In *Proceedings of Thirty-First AAAI Conference on Artificial Intelligence 2017*, 3351-3357.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., …Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of International conference on machine learning 2015*, 2048-2057.

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. In *Proceedings of IJCAI 2017*, 3901-3907. doi: /10.24963/ijcai.2017/545

Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., … Qiao, Y. (2018). A Multi-task Learning Approach for Image Captioning. In *Proceedings of IJCAI 2018*, 1205-1211. doi: 10.24963/ijcai.2018/168